The Korean Journal of Applied Statistics
2024, Vol. 37, No. 2, 225–237

DOI: https://doi.org/10.5351/KJAS.2024.37.2.225
Print ISSN1225-066X / Online ISSN 2383-5818

Analysis of the cause-specific proportional hazards model with missing covariates

Minjung Lee[1,a]

[a]Department of Statistics, Kangwon National University

Abstract

In the analysis of competing risks data, some of covariates may not be fully observed for some subjects. In such cases, excluding subjects with missing covariate values from the analysis may result in biased estimates and loss of efficiency. In this paper, we studied multiple imputation and the augmented inverse probability weighting method for regression parameter estimation in the cause-specific proportional hazards model with missing covariates. The performance of estimators obtained from multiple imputation and the augmented inverse probability weighting method is evaluated by simulation studies, which show that those methods perform well. Multiple imputation and the augmented inverse probability weighting method were applied to investigate significant risk factors for the risk of death from breast cancer and from other causes for breast cancer data with missing values for tumor size obtained from the Prostate, Lung, Colorectal, and Ovarian Cancer Screen Trial Study. Under the cause-specific proportional hazards model, the methods show that race, marital status, stage, grade, and tumor size are significant risk factors for breast cancer mortality, and stage has the greatest effect on increasing the risk of breast cancer death. Age at diagnosis and tumor size have significant effects on increasing the risk of other-cause death.

Keywords: augmented inverse probability weighted method, cause-specific proportional hazards model, competing risks, missing covariates, multiple imputation

1. 서론

경쟁위험자료에서 일부 공변량은 항상 관측되지만 일부 공변량은 연구 대상들의 일부분에 대해서만 관측될 수 있다. 결측된 공변량을 가진 경쟁위험자료는 암 연구 또는 의학 연구에서 흔히 관측되며, 이러한 자료를 분석하기 위한 통계 방법론의 연구가 필요하다. 결측된 공변량을 처리하기 위한 가장 간단한 방법은 결측된 공변량을 가진 연구 대상들을 분석에서 제외하고 공변량이 완벽히 관측된 자료만을 이용하여 분석하는 것이다. 그러나 이러한 방식은 편향된 결과를 도출할 뿐만 아니라 추정치의 효율성 손실을 초래한다. 따라서 결측된 공변량 자료를 분석에 포함시키고 자료를 분석하기 위한 방법을 연구하는 것이 필요하다. 본 연구에서는 경쟁위험자료의 분석에서 일부 공변량들이 결측값을 가질 때, 원인별 비례위험모형의 추정 방법을 연구하고자 한다.

This study has been worked with the support of a research grant of Kangwon National University in 2022. The author thanks the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial (PLCO-1248).

[1]Corresponding author: Department of Statistics, Kangwon National University, 1 Kangwondaehakgil, Chuncheon-si, Gangwon state 24341, Korea. E-mail: mlee@kangwon.ac.kr

Published 30 April 2024 / journal homepage: http://kjas.or.kr

생존자료 분석에서 공변량이 결측값을 가질 때, 결측 문제를 해결하기 위한 여러 가지 통계 분석방법이 개발되어왔다. 그 중, 증대된 역 확률 가중 방법(augmented inverse probability weighted method) (Robins 등, 1994; Wang과 Chen, 2001; Qi 등, 2005)과 다중대체 방법(multiple imputation method) (Bartlett과 Taylor, 2016)이 가장 많이 사용되는 분석방법이다. 다중대체 방법은 자료의 특성에 맞게 여러 가지 통계 모형의 적용이 가능하고 Rubin (1987)의 공식에 의해 추정이 용이하다. 다중대체 방법은 공변량의 결측값을 대체하기 위해 대체모형(imputation model)을 설정하는 것이 필요하다. 증대된 역 확률 가중 방법은 선택 확률과 결측 공변량의 조건부 분포에 모형을 가정하는 것이 필요하며, 선택 확률과 결측 공변량의 조건부 분포에 대해 가정한 두 가지 모형 중 적어도 한 가지 모형만 정확하게 설정된다면 그 추정량은 일관성(consistent)을 가진다. 이를 이중 강건성(double robustness property)이라고 한다. 본 연구에서는 경쟁위험자료 분석에서 일부 공변량들이 결측값을 가질 때, 다중대체 방법과 증대된 역 확률 가중 방법을 이용하여 원인별 비례위험모형의 회귀모수와 분산을 추정하는 방법을 연구한다. 모의실험을 통해, 다중대체 방법에 의해 구해진 추정량, 증대된 역 확률 가중 방법에 의해 구해진 추정량, 결측된 공변량 자료를 제외하고 공변량이 완벽히 관측된 자료만을 이용하여 구해진 추정량을 비교한다. 결측된 공변량을 가진 경쟁위험자료의 분석에 세 가지 방법 (증대된 역 확률 가중 방법, 다중대체 방법, 완벽한 공변량 자료만 이용하는 방법)을 적용하여 그 결과를 비교하고 본 연구에서 연구한 추정방법들의 실용성을 설명하고 의미 있는 결론을 도출하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 공변량이 결측된 경쟁위험자료에 대해 다중대체 방법과 증대된 역 확률 가중 방법을 이용하여 원인별 비례위험모형의 회귀모수 및 분산 추정방법에 대해 설명한다. 3절에서는 2절에서 소개한 증대된 역 확률 가중 방법과 다중대체 방법에 의해 구해진 추정치의 성능을 평가하기 위해 모의실험을 수행한 결과를 보여준다. 4절에서는 결측된 공변량을 가진 유방암 자료에 대한 설명과 다중대체 방법과 증대된 역 확률 가중 방법을 원인별 비례위험모형에 적용하여 분석한 결과를 보여준다. 5절에서는 결론을 제시하며 마무리한다.

2. 연구 방법

2.1. 원인별 비례위험모형 추정

연구 시작 후 사건이 발생할 때까지 걸린 시간을 T, 우중도 절단시간(right censoring time)을 C, 사건의 원인을 ϵ이라고 하자. 본 논문에서는 사건의 원인이 두 가지인 경우를 고려하고, 관심 원인을 $\epsilon = 1$, 경쟁 원인을 $\epsilon = 2$로 표기한다. p개의 공변량들의 벡터를 Z라고 하자. 공변량 Z가 주어져 있을 때, 사건발생시간 T와 우중도 절단시간 C는 독립이라고 가정한다. 관측된 자료는 (X_i, δ_i, Z_i) $(i = 1, \ldots, n)$로 나타낼 수 있다. 여기서 $X = \min(T, C)$는 관측시간을 나타내며, $\delta = I(T \leq C)\epsilon$는 사건의 원인 또는 우중도 절단의 발생을 나타내는 지시자이며, $I(\cdot)$는 지시함수(indicator function)이다. 공변량 Z가 주어져 있을 때, 사건의 원인 k에 대한 원인별 위험함수는 다음과 같이 정의된다.

$$\lambda_k(t; Z) = \lim_{\Delta t \to 0} \Pr(t \leq T < t + \Delta t, \epsilon = k \mid T \geq t, Z)/\Delta t, \quad k = 1, 2.$$

원인별 위험함수에 공변량이 미치는 영향을 알아보기 위해 경쟁위험자료 분석에서 가장 많이 사용되는 모형은 각 원인별 위험함수에 비례위험모형 (Cox, 1972)을 가정하는 원인별 비례위험모형(cause-specific proportional hazards model) (Prentice 등, 1978)이다.

$$\lambda_k(t; Z) = \lambda_{0k}(t) \exp\left(\beta_k^T Z\right), \quad k = 1, 2, \tag{2.1}$$

여기서 $\lambda_{0k}(t)$는 사건의 원인 k에 대한 미지의 기저위험함수(unknown baseline hazard function)이며, β_k는 사건의 원인 k에 대한 $p \times 1$인 회귀모수(regression parameter)들의 벡터이다. 원인별 비례위험모형의 회귀모수 β_k

는 다음과 같은 추정방정식의 해로서 구해진다.

$$U(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i - \frac{S^{(1)}(\beta_k, t)}{S^{(0)}(\beta_k, t)} \right\} dN_{ki}(t) = 0, \quad k = 1, 2,$$

여기서 $S^{(r)}(\beta_k, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i^{\otimes r} \exp(\beta_k^T Z_i)$ $(r = 0, 1, 2)$, 열벡터 a에 대해 $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa^T$이며, $N_{ki}(t) = I(X_i \le t, \delta_i = k)$, $Y_i(t) = I(X_i \ge t)$, $\tau = \sup\{t : \Pr(Y(t) = 1) > 0\}$이다. Andersen과 Gill (1982)에 의한 증명과 유사한 증명 방법을 사용하여 $\hat{\beta}_k$가 β_k의 일치 추정량이며, $n^{1/2}(\hat{\beta}_k - \beta_k)$가 평균이 0이고 분산이 Σ_k^{-1}인 정규분포로 수렴함을 보일 수 있다. 여기서 $\Sigma_k = E(M_{\bar{Z}}^{\otimes 2})$이고, $M_{\bar{Z}} = \int_0^{\tau} \{Z - s^{(1)}(\beta_k, t)/s^{(0)}(\beta_k, t)\} dM_k(t)$, $s^{(r)}(\beta_k, t) = E\{S^{(r)}(\beta_k, t)\}$ $(r = 0, 1, 2)$, $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_i(u) \exp(\beta_k^T Z_i) \lambda_{0k}(u) du$이다.

2.2. 공변량이 결측된 경쟁위험자료

일부 공변량들이 연구대상들의 일부분에 대해 관측되지 않을 때, 이를 표기하기 위해 공변량 Z를 $Z = (Z_1, Z_2)$로 나누자. 여기서 Z_1는 모든 연구대상들에 대해 관측되지만, Z_2는 연구대상들의 일부분에 대해서만 관측된다. 공변량 Z_2가 결측 또는 관측됨을 나타내는 지시자를 R이라고 정의하면, 공변량 Z_2의 값이 관측되면 $R = 1$이고, 결측되면 $R = 0$이다. 결측된 공변량을 가지는 경쟁위험자료는 $(X_i, \delta_i, Z_{1i}, R_i Z_{2i}, R_i)$ $(i = 1, \ldots, n)$로 나타낼 수 있다.

본 논문에서는 공변량 Z_2가 무작위 결측(missing at random)이라고 가정한다 (Rubin, 1976). 무작위 결측의 뜻은 (X, δ, Z_1, Z_2)가 주어져 있을 때 공변량 Z_2가 관측될 조건부 확률은 관측된 자료 (X, δ, Z_1)에만 의존한다는 것을 의미한다. 즉, $\Pr(R = 1 \mid X, \delta, Z_1, Z_2) = \Pr(R = 1 \mid X, \delta, Z_1)$이다. 여기서 $\Pr(R = 1 \mid X, \delta, Z_1) = \pi(X, \delta, Z_1)$를 선택 확률이라고 부른다.

결측된 공변량이 있을 때, 원인별 비례위험모형을 추정하기 위해 Bartlett과 Taylor (2016)의 다중대체 방법과 증대된 역 확률 가중 방법 중 Qi 등 (2005)이 제안한 완전 증강 가중 방법(fully augmented weighted method)이 사용될 수 있다. 다음 절에서는 두 가지 방법을 소개하고, 다중대체 방법과 완전 증강 가중 방법을 이용하여 원인별 비례위험모형의 회귀모수 및 분산을 추정하는 방법을 설명한다.

2.3. 다중대체 방법

결측된 공변량 Z_2에 대해 무작위 결측을 가정하고, 원인별 비례위험모형을 가정한다. Bartlett과 Taylor (2016)의 다중대체 방법에서, $f(X, \delta \mid Z_1, Z_2, \psi)$를 모수 ψ를 가진 실질적인 모형(substantive model)이라고 명칭하고, 실질적인 모형이 정확하게 설정되었다고 가정한다. 즉, $f(X, \delta \mid Z_1, Z_2) = f(X, \delta \mid Z_1, Z_2, \psi)$이다. 본 연구에서 실질적인 모형은 원인별 비례위험모형이다. 공변량 Z_2에서의 결측값을 대체하기 위해, $f(Z_2 \mid X, \delta, Z_1)$에 모수 ω를 가진 모수적 모형 $f(Z_2 \mid X, \delta, Z_1, \omega)$을 지정한다. 대체모형이 정확하게 설정될 때, 다중대체 방법에 의해 구해진 추정량은 일치 추정량이 된다. 대체모형이 실질적인 모형과 호환이 되지 않으면 대체모형이 잘못 설정될 수도 있다. 대체모형이 실질적인 모형과 호환이 됨을 확인하기 위해, $f(Z_2 \mid X, \delta, Z_1) \propto f(X, \delta \mid Z_1, Z_2) f(Z_2 \mid Z_1)$임을 주목하자. $f(X, \delta \mid Z_1, Z_2)$에 실질적인 모형을 설정함과 함께, $f(Z_2 \mid Z_1)$에도 모수 ϕ를 가진 모형 $f(Z_2 \mid Z_1, \phi)$을 지정한다. 공변량 Z_2가 일변량 변수인 경우, $f(Z_2 \mid Z_1)$에 대한 모형은 공변량 Z_2의 타입에 따라 적절한 모형으로 선택한다. 예를 들어, 연속형 변수에 대해서는 선형회귀모형, 이진 변수에 대해서는 로지스틱 회귀모형, 범주형 변수에 대해서는 다범주 로지스틱 회귀모형 등이 사용될 수 있다. 카운트 변수에 대해서는 포아송 또는 음이항 모형이 사용될 수 있다. ψ와 ϕ가 주어져 있을 때 $f(Z_2 \mid X, \delta, Z_1, \psi, \phi) \propto f(X, \delta \mid Z_1, Z_2, \psi) f(Z_2 \mid Z_1, \phi)$이므로, 공변량 Z_2의 결측값은 $f(X, \delta \mid Z_1, Z_2, \psi) f(Z_2 \mid Z_1, \phi)$에 비례하는 분포로부터 대체한다. 실제 자료 분석에서, Z_2는 여러 결측 패턴이 있는 벡터 값일 수도 있다 (즉, $Z_2 = (Z_{21}, \ldots, Z_{2q})$). 이런 경우에는 $f(Z_2 \mid Z_1)$에 모형을 지정하는 대신, 부분적으로 관측된 공변량 Z_{2j}에 대해

모형 $f(Z_{2j} \mid Z_{-2j}, Z_1, \phi_j)$을 지정한다 ($j = 1, \ldots, q$). 여기서 Z_{-2j}는 j 번째 원소를 제외한 Z_2의 나머지 원소들을 나타낸다. 공변량 Z_2가 벡터인 경우, Z_{2j}의 대체모형은 다음과 같이 실질적인 모형과 호환됨을 확인할 수 있다: $f(Z_{2j} \mid Z_{-2j}, X, \delta, Z_1, \psi, \phi_j) \propto f(X, \delta \mid Z_1, Z_2, \psi) f(Z_{2j} \mid Z_{-2j}, Z_1, \phi_j)$. ψ와 ϕ_j가 주어져 있을 때 Z_{2j}의 결측값은 $f(X, \delta \mid Z_1, Z_2, \psi) f(Z_{2j} \mid Z_{-2j}, Z_1, \phi_j)$에 비례하는 분포로부터 대체한다.

다중대체 방법은 관측된 자료가 주어져 있을 때 공변량의 결측값을 결측 자료의 사후분포로부터 랜덤하게 선택된 값으로 대체하는 것으로 구성되며, 사후분포는 베이지안 모형을 기반으로 한다. 대부분의 경우 사후분포들은 닫힌 형태가 아니기 때문에, 사후분포로터 자료를 추출하기 위해 Gibbs 샘플링을 이용한다. 결측된 공변량 Z_2가 일변량 변수인 경우, 각 원인별 비례위험모형의 모수와 ϕ에 대해 독립적인 사전분포를 가정하면서 Gibbs 샘플러는 공변량 Z_2의 결측값을 대체하기 위해 각 반복에서 다음 분포로부터 자료를 추출한다:

$$
\begin{aligned}
&f\left(\beta_k, \Lambda_{0k}(\cdot) \mid X, \delta, Z_1, Z_2^{obs}, Z_2^{imp}\right), \quad k = 1, 2 \\
&f\left(\phi \mid Z_1, Z_2^{obs}, Z_2^{imp}\right) \\
&f\left(Z_2^{imp} \mid X, \delta, Z_1, \phi, \beta_1, \Lambda_{01}(\cdot), \beta_2, \Lambda_{02}(\cdot)\right),
\end{aligned}
$$

여기서 Z_2^{obs}과 Z_2^{imp}는 모든 연구대상들에 대해 공변량 Z_2의 관측된 값과 Z_2의 현재 대체된 값을 나타낸다. 다중대체 알고리즘은 모든 모형의 모수에 대해 비정보적(non-informative) 사전분포를 사용한다. 공변량 Z_2가 벡터인 경우, 각 원인별 비례위험모형의 모수와 ϕ_j에 대해 독립적인 사전분포를 가정하면서 Gibbs 샘플러는 Z_{2j}의 결측값을 대체하기 위해 각 반복에서 다음 분포로부터 순차적으로 자료를 추출한다:

$$
\begin{aligned}
&f\left(\beta_k, \Lambda_{0k}(\cdot) \mid X, \delta, Z_1, Z_2^{obs}, Z_2^{imp}\right), \quad k = 1, 2 \\
&f\left(\phi_j \mid Z_1, Z_2^{obs}, Z_2^{imp}\right) \\
&f\left(Z_{2j}^{imp} \mid X, \delta, Z_1, Z_{-2j}, \phi_j, \beta_1, \Lambda_{01}(\cdot), \beta_2, \Lambda_{02}(\cdot)\right).
\end{aligned}
$$

M번 대체된 자료를 생성한 후, 원인별 비례위험모형의 회귀모수 β_k는 각 대체된 자료로부터 추정된다. 즉, M번 대체된 자료에 원인별 비례위험모형을 적용하여 구한 사건 원인 k의 회귀모수 추정량들은 $\hat{\beta}_k^1, \ldots, \hat{\beta}_k^M$이고, 그 분산 추정량들은 $\widehat{\text{var}}(\hat{\beta}_k^1), \ldots, \widehat{\text{var}}(\hat{\beta}_k^M)$이다 ($k = 1, 2$). 최종 회귀모수 추정량과 분산 추정량은 다음과 같이 Rubin (1987)의 공식에 의해 구해진다:

$$
\hat{\beta}_k^{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_k^m, \quad \widehat{\text{var}}\left(\hat{\beta}_k^{MI}\right) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\text{var}}\left(\hat{\beta}_k^m\right) + (1 + 1/M) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}_k^m - \hat{\beta}_k^{MI}\right)^2.
$$

2.4. 증대된 역 확률 가중 방법

경쟁위험이 존재하지 않는 생존자료에서 일부 공변량들이 결측값을 가질 때, 비례위험모형 (Cox, 1972)의 회귀모수를 추정하기 위하여 Qi 등 (2005)은 단순 가중 방법(simple weighted method)과 완전 증강 가중 방법(fully augmented weighted method)을 제안하였다. 경쟁위험이 존재하는 경우, 원인별 비례위험모형 (2.1)의 회귀모수 β_k를 추정하기 위해 단순 가중 방법의 추정방정식은 다음과 같이 확장할 수 있으며, 단순 가중 방법에 의한 β_k의 추정량은 아래 추정방정식을 만족하는 해로서 구해진다.

$$
U_{sw}(\beta_k, \pi) = \sum_{i=1}^{n} \frac{R_i}{\pi_i} \int_0^{\tau} \left\{ Z_i - \frac{S_{sw}^{(1)}(\beta_k, \pi, t)}{S_{sw}^{(0)}(\beta_k, \pi, t)} \right\} dN_{ki}(t) = 0, \quad k = 1, 2,
$$

여기서 $\pi = \Pr(R = 1 \mid X, \delta, Z_1)$는 선택 확률이며, $S_{sw}^{(r)}(\beta_k, \pi, t)$ $(r = 0, 1)$는 다음과 같다.

$$S_{sw}^{(r)}(\beta_k, \pi, t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\pi_i}Y_i(t)Z_i^{\otimes r}\exp\left(\beta_k^T Z_i\right)\right\}.$$

단순 가중 방법의 추정방정식은 공변량이 완벽히 관측된 자료에 선택 확률의 역수를 가중치로 사용한다. 따라서 선택 확률에 가정한 모형이 정확하게 설정되면 단순 가중 방법에 의해 도출된 추정량은 일관성을 가진다. 그러나 선택 확률에 가정한 모형이 정확하게 설정되지 않으면 그 추정량의 일관성은 영향을 받을 수 있으며, 공변량이 결측된 자료를 추정방정식에 사용하기 않기 때문에 그 추정량의 효율성이 떨어질 수 있다. 추정량의 효율성을 높이기 위해 공변량이 완벽히 관측된 자료 뿐만 아니라 공변량이 결측된 자료 모두 사용하는 완전 증강 가중 방법이 제안되었다. 경쟁위험이 존재할 때, 원인별 비례위험모형 (2.1)의 회귀모수 β_k를 추정하기 위해 완전 증강 가중 방법의 추정방정식은 다음과 같이 확장할 수 있다.

$$\begin{aligned} U_{faw}(\beta_k, \pi) = &\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau}\left\{Z_i - \frac{S_{aw}^{(1)}(\beta_k, \pi, E, t)}{S_{aw}^{(0)}(\beta_k, \pi, E, t)}\right\}dN_{ki}(t) \\ &+ \sum_{i=1}^{n}\left(1 - \frac{R_i}{\pi_i}\right)\int_0^{\tau}\left\{E(Z_i \mid X_i, \delta_i, Z_{1i}) - \frac{S_{aw}^{(1)}(\beta_k, \pi, E, t)}{S_{aw}^{(0)}(\beta_k, \pi, E, t)}\right\}dN_{ki}(t) = 0, \end{aligned} \tag{2.2}$$

여기서 (2.2)식의 첫 번째 항은 선택 확률의 역수를 가중치로 사용하며, 두 번째 항은 결측된 공변량 자료를 사용하는 증강항(augmentation term)을 나타낸다. 위의 추정방정식에서 $S_{aw}^{(r)}(\beta_k, \pi, E, t)$ $(r = 0, 1)$는 다음과 같다.

$$S_{aw}^{(r)}(\beta_k, \pi, E, t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\pi_i}Y_i(t)Z_i^{\otimes r}\exp\left(\beta_k^T Z_i\right) + \left(1 - \frac{R_i}{\pi_i}\right)Y_i(t)E\left(Z_i^{\otimes r}\exp\left(\beta_k^T Z_i\right) \mid X_i, \delta_i, Z_{1i}\right)\right\},$$

여기서 E는 미지의 결측 공변량의 조건부 기대값을 나타내며, $S_{aw}^{(r)}(\beta_k, \pi, E, t)$ $(r = 0, 1)$는 공변량이 완벽히 관측된 자료와 결측된 자료 모두 사용한다. 완전 증강 가중 방법은 추정방정식에서 증강항을 사용한다는 점에서 단순 가중 방법과 차이점이 있다. 완전 증강 가중 방법은 추정방정식 (2.2)의 두 번째 항과 $S_{aw}^{(r)}(\beta_k, \pi, E, t)$ $(r = 0, 1)$의 두 번째 항 모두에서 증강항을 사용하며, 완전 증강 가중 방법은 증강항을 통해 공변량이 결측된 불완전한 자료를 활용하기 때문에 완전 증강 가중 방법에 의해 도출된 추정량은 단순 가중 방법에 의해 도출된 추정량보다 더 높은 효율성을 가진다. 또한, 그 추정량은 이중 강건성을 가진다. 즉, 선택 확률과 결측된 공변량의 조건부 분포에 가정한 두 가지 모형 중 적어도 한 가지 모형이 정확하게 설정되면 그 추정량은 일관성을 가진다 (Wang과 Chen, 2001; Qi 등, 2005).

추정방정식 (2.2)에서 선택 확률 π_i는 주로 알려져 있지 않으며, 관측된 자료를 기반으로 추정해야 한다. 모수적 가정에 대한 의존도를 완화시키기 위해, Qi 등 (2005)은 관측된 자료를 기반으로 비모수적으로 선택 확률을 추정하는 방법을 제안하였다. 선택 확률의 추정에 사용되는 변수들을 $W = (X, \delta, Z_1)$라고 하자. W에 연속형 변수가 포함된 경우, 다음과 같은 Nadaraya-Watson 추정량을 이용하여 (Nadaraya, 1964; Watson, 1964) 선택 확률 $\pi(w)$를 추정한다.

$$\hat{\pi}(w) = \frac{\sum_{i=1}^{n} R_i K_h(w - W_i)}{\sum_{i=1}^{n} K_h(w - W_i)}, \tag{2.3}$$

여기서 d개의 연속형 변수가 W에 포함되어 있다면, $u = (u_1, u_2, \ldots, u_d) \in R^d$에 대해 $K_h(u) = (1/h^d)\prod_{i=1}^{d} K(u_i/h)$ 이며, $K(u)$는 q차 $(q > d)$의 커널 함수이고, h는 평활모수이다. q차 커널 함수 $K(u)$는 $\int K(u)du = 1$, $\int u^m K(u)du = 0$ $(m = 1, \ldots, (q-1))$, $\int u^q K(u)du \neq 0$, $\int K(u)^2 du < \infty$를 만족하는 평활함수이며, h는 $n \to \infty$에 따라 $nh^{2d} \to \infty$와 $nh^{2q} \to 0$를 만족하는 평활모수이다. Wang과 Wang (2001)과 Qi 등 (2005)는 $p > 2d$를 만족하는 정수 p

와 $q \geq p - d$를 만족하는 가장 작은 짝수 q에 대해 평활모수를 $h = O(n^{-1/p})$로 선택하였다. W가 연속형 변수와 이산형 변수 모두 포함하는 경우, 이산형 변수에 따라 층화한 다음 각 층에서 Nadaraya-Watson 추정량을 계산하여 선택 확률 $\pi(w)$를 추정한다.

추정방정식 (2.2)와 $S_{aw}^{(r)}(\beta_k, \pi, E, t)$에서 조건부 기대값 $E(Z_i \mid W_i)$와 $E(Z_i^{\otimes r} \exp(\beta_k^T Z_i)|W_i)$는 공변량이 완벽히 관측된 자료를 기반으로 다음과 같이 Nadaraya-Watson 추정량을 이용하여 추정한다.

$$\hat{E}\left(Z_i \mid W_i = w\right) = \frac{\sum_{j=1}^{n} Z_j R_j K_h\left(w - W_j\right)}{\sum_{j=1}^{n} R_j K_h\left(w - W_j\right)},$$

$$\hat{E}\left(Z_i^{\otimes r} \exp\left(\beta_k^T Z_i\right) \mid W_i = w\right) = \frac{\sum_{j=1}^{n} Z_j^{\otimes r} \exp\left(\beta_k^T Z_j\right) R_j K_h\left(w - W_j\right)}{\sum_{j=1}^{n} R_j K_h\left(w - W_j\right)},$$

여기서 모형 (2.1)에는 포함되지 않지만 결측된 공변량 Z_2와 상관관계가 있는 변수 A가 W에 포함되어 위의 조건부 기대값 추정에 사용될 수 있다. 선택 확률과 조건부 기대값의 비모수적 추정을 위해 서로 다른 커널 함수를 사용할 수 있으나, 단순화를 위해 Qi 등 (2005)는 동일한 커널 함수를 사용하였다.

결측 공변량의 조건부 기대값을 Nadaraya-Watson 추정량으로 대체한 추정방정식 $\hat{U}_{faw}(\beta_k, \pi) = 0$의 해를 $\hat{\beta}_k(\pi, \hat{E})$라고 하고, 선택 확률 $\pi(w)$를 추정 확률 $\hat{\pi}(w)$로 대체한 추정방정식 $\hat{U}_{faw}(\beta_k, \hat{\pi}) = 0$의 해를 $\hat{\beta}_k(\hat{\pi}, \hat{E})$라고 하자. $U_{faw}(\beta_k, \pi) = 0$와 $U_{faw}(\beta_k, \hat{\pi}) = 0$의 각각의 해를 $\hat{\beta}_k(\pi, E), \hat{\beta}_k(\hat{\pi}, E)$라고 하자. Appendix A의 규칙성 조건 (a1)–(a11) 아래, Qi 등 (2005)의 증명과 유사한 증명 방법을 사용하여 $\hat{\beta}_k(\pi, E), \hat{\beta}_k(\hat{\pi}, E), \hat{\beta}_k(\pi, \hat{E}), \hat{\beta}_k(\hat{\pi}, \hat{E})$ 모두 β_k의 일치 추정량이며, 추정량 모두 평균 0, 분산이 $\Sigma_k^{-1}\Sigma_k(\pi)\Sigma_k^{-1}$인 정규 분포로 수렴한다는 것을 보일 수 있다. 여기서 $\Sigma_k(\pi) = \Sigma_k + \Sigma_k^*(\pi)$, $\Sigma_k^*(\pi) = E\{(\pi^{-1} - 1)\mathrm{var}(M_{\tilde{Z}} \mid W)\}$이다. 이는 완전 증강 가중 추정량의 점근적 성질이 선택 확률과 결측 공변량의 조건부 기대값의 비모수적 추정에 영향을 받지 않는다는 것을 나타낸다. $\hat{\beta}_k$의 분산은 $\hat{\Sigma}_k^{-1}\hat{\Sigma}_k(\pi)\hat{\Sigma}_k^{-1}$로 추정된다. 여기서

$$\hat{\Sigma}_k = -\frac{1}{n}\frac{\partial}{\partial \beta_k}\hat{U}_{faw}\left(\hat{\beta}_k\left(\hat{\pi}, \hat{E}\right), \hat{\pi}\right), \quad \hat{\Sigma}_k^*(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i\left(1 - \hat{\pi}_i\right)}{\hat{\pi}_i^2}\left(\hat{M}_{\tilde{Z},i} - \hat{M}_{\tilde{Z},i}^0\right)^{\otimes 2},$$

$\hat{\Sigma}_k(\pi) = \hat{\Sigma}_k + \hat{\Sigma}_k^*(\pi)$이며, $\hat{M}_{\tilde{Z},i}$는 $M_{\tilde{Z},i}$의 추정량이며 다음과 같이 주어진다.

$$\hat{M}_{\tilde{Z},i} = \int_0^{\tau}\left[Z_i - \frac{S_{aw}^{(1)}\left(\hat{\beta}_k\left(\hat{\pi}, \hat{E}\right), \hat{\pi}, \hat{E}, t\right)}{S_{aw}^{(0)}\left(\hat{\beta}_k\left(\hat{\pi}, \hat{E}\right), \hat{\pi}, \hat{E}, t\right)}\right]\left[dN_{ki}(t) - Y_i(t)\exp\left(\hat{\beta}_k\left(\hat{\pi}, \hat{E}\right)^T Z_i\right)d\hat{\Lambda}_{0k}\left(t, \hat{\pi}\right)\right],$$

$$d\hat{\Lambda}_{0k}\left(t, \hat{\pi}\right) = \frac{1}{n}\frac{\sum_{i=1}^{n} dN_{ki}(t)}{S_{aw}^{(0)}\left(\hat{\beta}_k\left(\hat{\pi}, \hat{E}\right), \hat{\pi}, \hat{E}, t\right)},$$

$\hat{M}_{\tilde{Z},i}^0$은 $M_{\tilde{Z},i}^0 = E(M_{\tilde{Z},i} \mid W_i)$의 추정량이며 Nadaraya-Watson 추정량을 이용하여 구해진다.

3. 모의실험

공변량이 결측된 경쟁위험자료에서, 2절에서 소개한 다중대체 방법과 완전 증강 가중 방법을 적용하여 구한 원인별 비례위험모형의 회귀모수 추정량의 성능을 평가하고, 이를 전체 코호트(full cohort)에서 구한 회귀모수 추정량과 결측된 공변량 자료를 제외하고 공변량이 완벽히 관측된 자료만을 이용하여 구한 회귀모수 추정량과의 성능을 비교하기 위해 모의실험을 수행하였다.

두 개의 공변량 Z_1과 Z_2를 생성하였고, 공변량 Z_1은 표준정규분포에서, 공변량 Z_2는 성공 확률이 0.5인 베르누이 분포에서 생성하였다. 공변량 Z_1은 모든 연구대상들에 대해 관측되지만, 공변량 Z_2는 연구대상 일부분에 대해서만 관측된다. 결측 공변량 Z_2와 상관계수가 0.85이고 성공 확률이 0.5인 변수 A를 베르누이 분포에서 생성하였다. 각 원인별 기저위험함수에 지수분포와 콤펄즈 분포의 위험함수를 가정하였고, 각 원인별

Table 1: Simulation results for the regression paramter estimates, where $Z_1 \sim N(0,1)$, $Z_2 \sim B(1,0.5)$, $\pi(X,\delta,Z_1) = (1+\exp(2.76-\delta-X-Z_1))^{-1}$, 39.7% missing covariate data

$n = 200$	$\hat{\beta}_{11}$				$\hat{\beta}_{12}$			
	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP
Full cohort	0.0075	0.018	0.016	0.938	−0.0002	0.061	0.060	0.946
CC analysis	0.2709	0.037	0.033	0.673	0.0360	0.117	0.112	0.948
MI method	0.0096	0.018	0.016	0.943	−0.0114	0.103	0.106	0.953
FAW-$\hat{\pi}(X,\delta)$	0.0086	0.018	0.020	0.953	−0.0065	0.081	0.075	0.935
FAW-$\hat{\pi}(X,\delta,Z_1)$	0.0081	0.018	0.019	0.950	−0.0079	0.080	0.072	0.933
	$\hat{\beta}_{21}$				$\hat{\beta}_{22}$			
	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP
Full cohort	0.0155	0.020	0.018	0.942	0.0069	0.064	0.066	0.959
CC analysis	0.1122	0.028	0.025	0.884	0.0472	0.088	0.084	0.942
MI method	0.0169	0.020	0.019	0.943	0.0162	0.086	0.085	0.953
FAW-$\hat{\pi}(X,\delta)$	0.0166	0.020	0.021	0.951	0.0154	0.072	0.074	0.959
FAW-$\hat{\pi}(X,\delta,Z_1)$	0.0167	0.020	0.020	0.951	0.0161	0.072	0.074	0.955
$n = 500$	$\hat{\beta}_{11}$				$\hat{\beta}_{12}$			
	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP
Full cohort	0.0028	0.007	0.006	0.938	0.0047	0.023	0.023	0.948
CC analysis	0.2455	0.014	0.012	0.398	0.0501	0.042	0.043	0.951
MI method	0.0037	0.007	0.006	0.938	0.0071	0.038	0.041	0.958
FAW-$\hat{\pi}(X,\delta)$	0.0032	0.007	0.008	0.961	0.0057	0.031	0.030	0.944
FAW-$\hat{\pi}(X,\delta,Z_1)$	0.0031	0.007	0.007	0.961	0.0045	0.031	0.029	0.939
	$\hat{\beta}_{21}$				$\hat{\beta}_{22}$			
	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP	Bias	Emp.Var	E($\widehat{\text{var}}$)	CP
Full cohort	0.0001	0.007	0.007	0.955	0.0143	0.027	0.025	0.943
CC analysis	0.0840	0.009	0.009	0.863	0.0466	0.035	0.032	0.928
MI method	0.0006	0.007	0.007	0.951	0.0174	0.034	0.032	0.946
FAW-$\hat{\pi}(X,\delta)$	0.0005	0.007	0.008	0.956	0.0176	0.030	0.028	0.944
FAW-$\hat{\pi}(X,\delta,Z_1)$	0.0006	0.007	0.007	0.957	0.0168	0.030	0.028	0.941

위험함수는 $\lambda_1(t;Z_1,Z_2) = \alpha_1 \exp(\beta_{11}Z_1+\beta_{12}Z_2)$, $\lambda_2(t;Z_1,Z_2) = (\alpha_2/\theta_2)(t/\theta_2)^{\alpha_2-1}\exp(\beta_{21}Z_1+\beta_{22}Z_2)$이다. 여기서 $(\alpha_1,\alpha_2,\theta_2,\beta_{11},\beta_{12},\beta_{21},\beta_{22}) = (0.1,-3,0.3,0.5,0.6,0.3,0.4)$이다. 사건발생시간 T는 $1-\exp\{-\int_0^t \lambda_1(u;Z_1,Z_2)du - \int_0^t \lambda_2(u;Z_1,Z_2)du\}$에서 생성하였고, 사건 원인 ϵ은 성공 확률이 $\lambda_1(T;Z_1,Z_2)/\{\lambda_1(T;Z_1,Z_2)+\lambda_2(T;Z_1,Z_2)\}$인 베르누이 분포에서 생성하였다. 중도절단시간 C는 균일 분포 $U(0,\gamma_c)$에서 생성하였으며, γ_c는 중도절단 비율이 30%가 되도록 선택하였다. 관측시간은 $X = \min(T,C)$이고, $\delta = I(T \le C)\epsilon$이다. 결측 지시자 R은 성공확률이 $\pi(X,\delta,Z_1) = (1+\exp(2.76-\delta-X-Z_1))^{-1}$인 베르누이 분포에서 생성하였고, $R=0$인 경우 공변량 Z_2는 결측값을 가진다. 본 실험에서, 39.7%의 결측 공변량 자료가 발생하였다. 자료 크기가 200과 500인 모의실험을 각각 1,000번 수행하였고, 다중대체 방법, 완전 증강 가중 방법, 전체 코호트, 공변량이 완벽히 관측된 자료만을 이용하는 방법으로부터 원인별 비례위험모형의 회귀모수 추정치와 분산 추정치를 구하였다.

다중대체 방법에서 $f(Z_2 \mid Z_1)$에 Z_1에 기반한 로지스틱 회귀모형을 지정하였고, 10번 대체하였다. 10번 대체된 각 자료에 원인별 비례위험모형을 적합하고 Rubin의 공식을 사용하여 회귀모수 추정치와 분산 추정치를 구하였다. 완전 증강 가중 방법에서 선택 확률 $\pi(W)$는 $W^* = (X,\delta)$ 또는 $W = (X,\delta,Z_1)$를 기반으로 정규 커널과 평활모수 $h = 4\sigma_W^* n^{-1/3}$ 또는 $h = 3\sigma_W n^{-1/5}$를 이용하여 Nadaraya-Watson 추정량에 의해 추정하였다. 여기서 $W^* = (X,\delta)$를 기반으로 $\pi(W)$를 추정했을 때는 σ_W^*는 δ로 층화한 X의 표본표준편차이고, $W = (X,\delta,Z_1)$

Table 2: Data characteristics

Categorical variables		n	%
Race	White	1381	90.80
	Black	58	3.81
	Other	82	5.39
Marital status	Married	1043	68.57
	Single	478	31.43
Stage	0	172	11.31
	I	842	55.36
	II	436	28.67
	III	60	3.94
	IV	11	0.72
Grade	Low	1111	73.04
	High	410	26.96
Tumor size	< 2 cm	1024	67.32
	≥ 2 cm	320	21.04
	Missing	177	11.64
Continuous variable		Mean	Standard deviation
Age at diagnosis		68.74	6.5

를 기반으로 $\pi(W)$를 추정했을 때는 σ_W는 δ로 층화한 X의 표본표준편차와 Z_1의 표본표준편차들의 벡터이다. 결측 공변량의 조건부 기대값은 $W^\dagger = (X, \delta, Z_1, A)$를 기반으로 정규 커널과 평활모수 $h = 3\sigma_W^\dagger n^{-1/5}$을 이용하여 Nadaraya-Watson 추정량에 의해 추정하였다. 여기서 $\sigma_W^\dagger$는 δ와 A로 층화한 X의 표본표준편차와 Z_1의 표본표준편차들의 벡터이다. $W = (X, \delta, Z_1)$와 $W^\dagger = (X, \delta, Z_1, A)$를 기반으로 선택 확률과 결측 공변량의 조건부 기대값을 추정한 경우는 선택 확률과 결측 공변량의 조건부 기대값 모두에 정확한 모형을 설정하여 추정한 경우에 해당하며, 결측 공변량의 조건부 기대값은 $W^\dagger = (X, \delta, Z_1, A)$를 기반으로 추정하였으나 선택 확률은 $W^* = (X, \delta)$를 기반으로 추정한 경우는 결측 공변량의 조건부 기대값에는 정확한 모형이 설정되었으나 선택 확률에는 정확한 모형이 설정되지 않은 경우에 해당한다. 본 실험에서는, 이 두 가지 경우를 통해 완전 증강 가중 방법에 의해 구해진 회귀모수 추정치의 이중 강건성을 평가하였다.

Table 1은 다중대체(MI) 방법, 완전 증강 가중(FAW) 방법, 전체 코호트(Full cohort), 공변량이 완벽히 관측된 자료(CC analysis)로부터 구한 원인별 비례위험모형의 회귀모수 추정치의 편향(Bias), 경험적 분산(Emp.Var), 분산 추정치들의 평균(E($\widehat{\text{var}}$)), 95% 신뢰구간에 대한 경험적 범위 확률(CP)을 보여준다. 다중대체 방법에 의해 구해진 회귀모수 추정치는 편향이 작았고, 분산 추정치는 경험적 분산과 일치하였고, 경험적 범위 확률은 명목 수준에 가까웠다. 완전 증강 가중 방법에 의해 구해진 회귀모수 추정치는 선택 확률 π가 일관되지 않게 추정되더라도 편향이 작았음을 보여주었고, 이로부터 이중 강건성을 확인할 수 있었다. 완전 증강 가중 방법에 의해 구해진 분산 추정치는 경험적 분산과 일치하였고, 경험적 범위 확률은 명목 수준에 가까웠다. 다중대체 방법과 완전 증강 가중 방법에 의해 구해진 각 사건 원인에 대한 Z_1의 회귀모수 추정치 $\hat{\beta}_{11}, \hat{\beta}_{21}$는 전체 코호트로부터 구해진 추정치 $\hat{\beta}_{11}, \hat{\beta}_{21}$의 효율성을 거의 달성하였다. 완전 증강 가중 방법에 의해 구해진 결측된 공변량 Z_2의 회귀모수 추정치 $\hat{\beta}_{12}, \hat{\beta}_{22}$는 다중대체 방법에 의해 구해진 추정치 $\hat{\beta}_{12}, \hat{\beta}_{22}$보다 작은 분산 추정치와 경험적 분산을 가졌다. 공변량이 완벽히 관측된 자료만을 이용하여 구한 회귀모수 추정치들은 큰 편향을 가졌고, 이로 인해 β_{11}, β_{21}의 95% 신뢰구간에 대한 경험적 범위 확률이 명목 수준보다 많이 낮았다. 또한, 전체 코호트 자료로부터 구해진 추정치와 비교하여 그 경험적 분산은 매우 크다. 이는 결측된 공변량 자료를 제외하고 분석함으로 인해 추정치에서 효율성 손실이 있었음을 나타낸다. 본 모의실험 결과는 다중대체 방법과 완전 증강 가중 방법에 의해 구해진 원인별 비례위험모형의 회귀모수 추정치들의 성능이 좋았음을

Table 3: Regression parameter estimates, standard errors, and p-values for death from breast cancer and for death from other causes under the cause-specific proportional hazards model

	Fully augmented weighted			Multiple imputation			Complete case		
Death from breast cancer	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	p-value	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	p-value	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	p-value
Race									
White	0	-	-	0	-	-	0	-	-
Black	0.946	0.291	0.001	0.965	0.276	< 0.001	0.920	0.285	0.001
Other	−0.233	0.392	0.552	−0.237	0.391	0.543	−0.365	0.459	0.426
Marital status									
Married	0	-	-	0	-	-	0	-	-
Single	0.454	0.170	0.007	0.442	0.169	0.009	0.461	0.171	0.007
Stage									
0-I	0	-	-	0	-	-	0	-	-
II	0.555	0.239	0.020	0.581	0.254	0.022	0.481	0.254	0.058
III-IV	2.134	0.271	< 0.001	2.173	0.282	< 0.001	2.172	0.287	< 0.001
Grade									
Low	0	-	-	0	-	-	0	-	-
High	0.511	0.167	0.002	0.508	0.167	0.002	0.517	0.172	0.003
Tumor size									
< 2 cm	0	-	-	0	-	-	0	-	-
≥ 2 cm	0.536	0.212	0.011	0.517	0.242	0.032	0.498	0.237	0.035
	Fully augmented weighted			Multiple imputation			Complete case		
Death from other causes	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	p-value	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	p-value	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	p-value
Age at diagnosis	0.118	0.008	< 0.001	0.118	0.007	< 0.001	0.118	0.008	< 0.001
Tumor size									
< 2 cm	0	-	-	0	-	-	0	-	-
≥ 2 cm	0.426	0.111	< 0.001	0.422	0.110	< 0.001	0.433	0.112	< 0.001

보여주었다.

4. 유방암 자료 분석 결과

4.1. 자료 설명

본 논문에서 분석한 자료는 미국 국립암연구소의 전립선, 폐, 대장, 난소(PLCO) 암 선별 시험 연구에서 제공받은 유방암 자료이다. 1993년부터 2001년까지 임상시험에 참여하고 2009년까지 추적 기간 동안 유방암 진단을 받은 여성들 중, 설문지를 작성하고 일상적인 치료 그룹(usual-care group)에 속하는 1521명에 대한 자료이다. 최대 관측시간은 9329개월이며, 1521명의 유방암 환자들 중, 154명이 유방암으로 사망하였고, 491명은 다른 원인으로 사망하였고, 나머지 876명은 중도절단 되었다. 분석에 사용된 공변량은 진단시 연령, 인종, 기혼여부, 병기, 분화도, 종양의 크기이다. 진단시 연령, 인종, 기혼여부, 병기, 분화도는 모든 환자들에게 관측되었지만, 종양의 크기는 177명의 환자들(11.64%)에게는 관측되지 않았다. Table 2은 분석에 사용된 공변량들의 기술통계를 보여준다.

4.2. 원인별 비례위험모형 적합 결과

진단시 연령, 인종, 기혼여부, 병기, 분화도, 종양의 크기가 유방암 사망 위험률과 다른 원인 사망 위험률에 미치는 영향을 파악하기 위해 원인별 비례위험모형을 적합하였다. 177명의 환자들에 대해 종양 크기가 관측되지 않았으므로, 원인별 비례위험모형에서 공변량들의 효과를 추정하기 위해 다중대체 방법과 완전 증강 가중 방법을 적용하였다. 다중대체 방법에서 진단시 연령, 인종, 기혼여부, 병기, 분화도를 기반으로 $f(Z_2 \mid Z_1)$에 로지스틱 회귀모형을 지정하였고, 5번 대체하였다. 완전 증강 가중 방법에서 선택 확률과 결측 공변량의 조건부 기대값은 W= (관측시간(X), 생존 상태(δ), 진단시 연령, 인종, 기혼여부, 병기, 분화도)를 기반으로 정규 커널과 평활모수 $h = 3\sigma_W n^{-1/5}$을 이용하여 Nadaraya-Watson 추정량을 이용하여 추정하였다. 여기서 σ_W는 생존 상태, 인종, 기혼여부, 병기, 분화도로 층화한 관측시간의 표본표준편차와 진단시 연령의 표본표준편차들의 벡터이다. 예비 분석에서 비례위험 가정이 각 사망 원인별 위험함수 모두에 적절함을 확인하였다.

Table 3는 종양의 크기가 관측되지 않은 177명의 환자들을 제외한 자료(complete case), 완전 증강 가중 방법, 다중대체 방법에 의해 구해진 사망 원인별 비례위험모형의 회귀모수 추정치와 그 표준오차를 보여준다. Table 3는 유의수준 5%에서 유의한 공변량만을 모형에 포함시키고 적합한 결과이다. 유의수준 5%에서, 모든 분석에서 인종, 기혼여부, 병기, 분화도, 종양의 크기는 유방암 사망 위험률에 유의한 영향을 미치는 요인들이였으나, 진단시 연령은 유의한 요인이 아니었다. 종양의 크기와 진단시 연령은 다른 원인 사망 위험률에 유의한 영향을 미치는 요인들이였으나, 인종, 기혼여부, 병기, 분화도는 유의한 요인이 아니었다. 미혼이고 병기와 분화도가 높을수록, 종양의 크기가 클수록 유방암으로 사망할 위험률이 높아지며, 진단시 연령이 높고 종양의 크기가 클수록 다른 원인으로 사망할 위험률이 높아짐을 알 수 있다. 완전 증강 가중 방법과 다중대체 방법에서 병기 II와 병기 0-I 사이에 유방암 사망 위험률에서 유의한 차이가 있었으나, 완벽한 공변량 자료만을 이용하는 분석에서는 두 병기 간에 유방암 사망 위험률에서 유의한 차이가 없었다. 이를 통해 공변량이 결측된 자료를 제외하고 완벽한 공변량 자료만을 이용하여 분석하면 편향된 추론을 얻을 수 있음을 확인할 수 있었다.

5. 결론

공변량이 결측된 자료는 경쟁위험분석에서 종종 관측되며, 본 논문에서는 결측된 공변량이 존재하는 경쟁위험자료에서 무작위 결측 가정 아래 다중대체 방법과 완전 증강 가중 방법에 의해 원인별 비례위험모형을 추정하는 방법을 연구하였다. 2절에서 소개한 다중대체 방법과 완전 증강 가중 방법은 공변량이 완벽히 관측된 자료 뿐만 아니라 공변량이 결측된 자료 모두 사용하며, 결측된 공변량 자료를 사용하는 방식은 다르다 (Qi 등, 2010). 완전 증강 가중 방법은 추정방정식에서 비모수적으로 공변량이 결측된 자료를 사용하며, 그 추정량은 이중 강건성을 가진다는 장점이 있다. 다중대체 방법은 공변량의 결측값을 대체하기 위해 대체된 자료를 생성할 때 공변량이 결측된 자료를 사용하며, 대체된 자료에 여러 가지 통계 모형의 적용이 가능하며 Rubin의 공식을 사용하여 추정량과 분산 추정량의 계산이 용이하다는 장점이 있다. 모의실험을 통해 다중대체 방법과 완전 증강 가중 방법에 의해 구해진 추정치들이 잘 작동하며, 공변량이 완벽히 관측된 자료만을 이용하는 분석과 비교하여 두 분석이 편향이 작고, 효율성이 높은 추정치들을 산출함을 확인하였다. 또한, 완벽히 관측된 공변량의 효과 추정시 그 추정치의 효율성이 개선됨을 확인하였다. 완전 증강 가중 방법의 경우, 이중 강건성으로 인해 선택 확률이 일관되지 않게 추정되어도 편향되지 않은 추정치를 도출할 수 있음을 모의실험을 통해 확인하였다.

미국 국립암연구소의 PLCO 암 선별 시험 연구에서 제공받은 유방암 자료에서 종양의 크기가 연구 대상들의 일부분에 대해서만 관측되었다. 유방암 사망 위험률에 유의한 영향을 미치는 공변량들을 파악하고 그 효과를 추정하기 위해 다중대체 방법과 완전 증강 가중 방법을 적용하여 원인별 비례위험모형을 적합한 결과, 인종, 기혼여부, 병기, 분화도, 종양의 크기는 유방암 사망 위험률을 높이는 유의한 요인들이였고, 종양의 크

기와 진단시 연령은 다른 원인 사망 위험률을 높이는 유의한 요인들이였다. 미혼이며 병기와 분화도가 높고, 종양의 크기가 클수록 유방암 사망 위험률이 높아지며, 고령이고 종양의 크기가 클수록 다른 원인 사망 위험률이 높아짐을 확인하였다. 모의실험과 유방암 자료 분석을 통하여 공변량이 결측된 자료를 제외하고 완벽한 공변량 자료만을 이용하여 분석하면 편향된 추론을 얻을 수 있음을 확인하였다. 본 논문에서 분석한 유방암 자료에는 수술과 항암치료에 관한 변수가 존재하지 않으나 이에 관한 변수를 모형에서 고려할 수 있으면 유방암 사망 위험률 예측에 더 유의미한 결론을 얻을 수 있을 것이라 생각된다. 본 논문에서 소개한 다중대체 방법과 완전 증강 가중 방법은 공변량이 결측된 다른 암 자료의 분석에도 활용가능하며 암 환자의 사망율 예측에 유의한 요인들을 파악하는데 도움이 될 것이라 기대한다.

본 논문에서는 공변량이 결측값을 가질 때, 다중대체 방법과 완전 증강 가중 방법을 이용하여 원인별 비례위험모형에서 공변량 효과를 추정하였다. 경쟁위험분석에서 원인별 비례위험모형 외에도 비례하위분포위험모형(proportional subdistribution hazards model) (Fine과 Gray, 1999)이 많이 사용되고 있으며, 이 방법은 누적발생함수(cumulative incidence function)에 대한 공변량의 효과를 직접 평가할 수 있다는 장점이 있다. 본 논문에서는 다중대체 방법과 완전 증강 가중 방법을 이용하여 결측된 공변량을 가진 원인별 비례위험모형의 추정을 연구하였으나, 비례하위분포위험모형을 사용하는 것을 고려해 볼 수 있다. 이는 본 연구의 향후 연구과제가 될 것이다.

Appendix A:

아래는 규칙성 조건이다.

(a1) $\Lambda_{0k}(\tau) < \infty$, $\Pr(Y(\tau) = 1) > 0$이다. 공변량 Z는 시간에 독립적이며 경계가 있다. 선택 확률 π는 0에서 멀리 떨어져 있다.

(a2) $v(\beta_k, t) = s^{(2)}(\beta_k, t)/s^{(0)}(\beta_k, t) - \left(s^{(1)}(\beta_k, t)/s^{(0)}(\beta_k, t)\right)^{\otimes 2}$라고 정의하자. $\Sigma_k = \int_0^\tau v(\beta_k, t)s^{(0)}(\beta_k, t)d\Lambda_{0k}(t)$는 양의 확정 (positive definite)이다.

(a3) $\eta > 0$에 대해, 선택 확률 $\pi \geq \eta$이다.

(a4) 선택 확률 $\pi(w)$는 거의 모든 곳에서 W의 연속 성분에 대해 q 연속 및 유계 편도함수를 가진다.

(a5) W의 확률밀도함수 $f(w)$와 $W|R$의 조건부 확률밀도함수 $f_{W|R}(w)$는 0에서 멀리 떨어져있다. $f(w)$ 및 $f_{W|R}(w)$는 거의 모든 곳에서 W의 연속 성분에 대해 q 연속 및 유계 편도함수를 갖는다.

(a6) 조건부 분포 $f_{W|R=0}(w)$ 및 $f_{W|R=1}(w)$는 동일한 지원을 가지며, $c(w) = f_{W|R=0}(w)/f_{W|R=1}(w)$는 지지대를 경계로 한다.

(a7) 선택 확률 $\pi(w)$와 조건부 기대값 $E\{Z^{\otimes r}\exp(\beta_k^T Z) \mid W = w\}$, $E\{(Z^{\otimes r}\exp(\beta_k^T Z))^{\otimes 2} \mid W = w\}$ $(r = 0, 1)$, $E\{M(\tau) \mid W = w\}$, $E\{M(\tau)^{\otimes 2} \mid W = w\}$, $E\{M_{\tilde{Z}} \mid W = w\}$, $E\{M_{\tilde{Z}}^{\otimes 2} \mid W = w\}$, $E\{M_Z \mid W = w\}$, $E\{M_Z^{\otimes 2} \mid W = w\}$는 거의 모든 곳에서 W의 연속 성분에 대해 q 연속 및 유계 편도함수를 가진다. 여기서 $M_Z = \int_0^\tau Z dM_k(t)$이다.

(a8) $n^{-1/2}\sum_{i=1}^n R_i\pi_i^{-2}(\hat{\pi}_i - \pi_i)M_{ki}(t)$ 는 연속 샘플 경로와 함께 평균 0인 가우시안 프로세스로 수렴한다.

(a9) $n \to \infty$에 따라 $\sup_{t\in[0,\tau]}\|n^{-1/2}\sum_{i=1}^n(1 - R_i/\pi_i)\int_0^\tau[\hat{E}\{dM_{ki}(u)|W_i\} - E\{dM_{ki}(u)|W_i\}]\| \xrightarrow{p} 0$이다.

(a10) $n^{-1/2}\sum_{i=1}^n(R_i(\hat{\pi}_i - \pi_i)/\pi_i^2)E\{M_{ki}(t)|W_i\}$는 연속 샘플 경로와 함께 평균 0인 가우시안 프로세스로 수렴한다.

(a11) $n \to \infty$에 따라 $\sup_{t\in[0,\tau]}\|n^{-1/2}\sum_{i=1}^n(1 - R_i/\hat{\pi}_i)\int_0^\tau[\hat{E}\{dM_{ki}(u)|W_i\} - E\{dM_{ki}(u)|W_i\}]\| \xrightarrow{p} 0$이다.

References

Andersen PK and Gill RD. (1982). Cox's regression model for counting processes: A large sample study, *Annals of Statistics*, **10**, 1100–1120.

Bartlett JW and Taylor JM. (2016). Missing covariates in competing risks analysis, *Biostatistics*, **17**, 751–763.

Cox DR. (1972). Regression models and life-tables (with discussion), *Journal of Royal Statistical Society, Series A*, **34**, 187–220.

Fine JP and Gray RJ. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–508.

Nadaraya EA. (1964). On estimating regression, *Theory of Probability and Its Applications*, **9**, 141–142.

Prentice RL, Kalbeisch JD, Peterson AV, Flournoy N, Farewell VT, and Breslow NE. (1978). The analysis of failure times in the presence of competing risks, *Biometrics*, **34**, 541–554.

Qi L, Wang CY, and Prentice RL. (2005). Weighted estimators for proportional hazards regression with missing covariates, *Journal of the American Statistical Association*, **472**, 1250–1263.

Qi L, Y-F Wang, and Y He. (2010). A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates, *Statistics in Medicine*, **29**, 2592–2604.

Robins JM, Rotnitzky A, and Zhao LP. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.

Rubin DB. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin DB. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Wang CY and Chen HY. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression, *Biometrics*, **57**, 414–419.

Wang S and Wang CY. (2001). A note on kernel assisted estimators in missing covariate regression, *Statistics & Probability Letters*, **55**, 439–449.

Watson GS. (1964). Smooth regression analysis, *Sankhya A*, **26**, 359–372.

Received August 22, 2023; Revised October 23, 2023; Accepted October 25, 2023

누락된 공변량을 가진 원인별 비례위험모형의 분석

이민정[1,a]

[a]강원대학교 통계학과

요 약

경쟁위험자료에서 일부 공변량들이 연구대상들의 일부분에 대해 관측되지 않을 수 있다. 그런 경우 결측된 공변량 값을 가진 연구대상들을 분석에서 제외하는 것은 편향된 추정치와 효율성 손실이 발생할 수 있다. 본 논문에서는 누락된 공변량을 가진 원인별 비례위험모형의 회귀모수 추정을 위해 다중대체 방법과 증대된 역 확률 가중 방법을 연구하였다. 모의실험을 통해 다중대체 방법과 증대된 역 확률 가중 방법에 의해 구해진 추정량의 성능을 평가한 결과, 이 방법들이 잘 수행됨을 확인하였다. 미국 국립암연구소의 전립선, 폐, 대장, 난소 암 선별 시험 연구에서 제공하는 종양 크기의 값이 누락된 유방암 자료에 대해 암 사망 위험률과 다른 원인 사망 위험률에 유의한 영향을 미치는 요인을 파악하기 위해 다중대체 방법과 증대된 역 확률 가중 방법을 적용하였다. 다중대체 방법과 증대된 역 확률 가중 방법에 의해 원인별 비례위험모형을 적합한 결과, 인종, 기혼여부, 병기, 분화도, 종양의 크기는 유방암 사망 위험률에 유의한 영향을 미치는 요인들이였으며, 병기가 유방암 사망 위험률을 높이는데 가장 큰 영향을 미치는 요인임을 확인하였다. 진단시 연령과 종양의 크기는 다른 원인 사망 위험률을 높이는데 유의한 영향을 미치는 요인이였다.

주요용어: 증대된 역 확률 가중 방법, 원인별 비례위험모형, 경쟁위험, 결측 공변량, 다중대체

이 논문은 2022년도 강원대학교 대학회계의 지원을 받아 수행한 연구임. 전립선, 폐, 대장, 난소(PLCO) 암 선별 시험에서 수집한 미국 국립암연구소의 자료에 접근할 수 있도록 승인해 준 미국 국립암연구소에 감사를 표함 (PLCO-1248).
[1]교신저자: (24341) 강원도 춘천시 강원대학길 1, 강원대학교 통계학과. E-mail: mlee@kangwon.ac.kr